Specification Amendments

Internet filtering products have previously been developed to attempt to filter Internet objects based on adult content. All filtering products require a method by which to classify Internet objects. The prior art methods of classification are detailed below, but can be summarized as taking one of three approaches: (i) filtering based on classification information embedded in an Internet object; (ii) compilation of "blacklists" and/or "whitelists" that filtering products may reference; and (iii) real-time textual analysis.

One Internet filtering product is produced by Netscape as part of the ~~Netscape~~ NETSCAPE web browser. The browser includes an adult content filtering feature that classifies web pages based on the PICS labeling system, a voluntary system by which Internet content providers include special codes in their Internet objects (in this case web pages) to describe the content of the object. The PICS labels are the only mechanism used by Netscape in classifying the adult content of web pages. The PICS system is described in greater detail at ~~http://~~www.w3.org/pics/. The drawback in relying solely on the PICS label to classify Internet objects is that not all sites are classified using the system, as participation in the PICS system is voluntary. In addition, there is no independent verification process, and users rely solely on the judgement of the Internet content providers, which may be biased by self-interest.

~~Cyber Patrol~~ CYBER PATROL is another Internet filtering product. ~~Cyber Patrol~~ CYBER PATROL maintains a "blacklist" of web sites that are considered to

contain adult content. The ~~Cyber Patrol~~ CYBER PATROL web page at ~~http://~~www.cyberpatrol.com/ discloses that "professional researchers compile" the lists, apparently manually. With the current growth rate of Internet users and Internet content providers, the current method of manual classification is inadequate.

~~SurfWatch~~ SURFWATCH [~~http://~~www1.surfwatch.com/] is another Internet filtering product that works by maintaining blacklists. ~~SurfWatch~~ SURFWATCH appears to search web pages and URLs for restricted keywords. Any page or URL containing a restricted keyword is classified as adult content. There is no further initial verification process. This can lead to a site being erroneously classified as adult, as illustrated in the recent incident in which one of the pages on the "Whitehouse for Kids" website was classified as adult content because it was named "couples.html".

~~CYBERsitter~~ CYBERSITTER is yet another Internet filtering product that attempts to classify web sites, by looking at the text of the page and URLS. The product removes profane English words from the text of web pages, but does not filter out pornographic images from web pages which do not contain text and does not filter out words that are profane words in foreign languages.

~~NetNanny~~ NETNANNY [~~http://~~www.netnanny.net/] is still another Internet filtering product that uses a blacklist of domain names of web sites not suitable for children. The ~~NetNanny~~ NETNANNY web site discloses that the

~~NetNanny~~ NETNANNY site list is compiled by ~~NetNanny~~ NETNANNY staff, the suggestions of customers, and third party children's advocacy groups. The ability of the product to filter out undesirable sites is limited by the comprehensiveness of the blacklist, which is compiled manually.

In sum, given the rapid proliferation of Internet objects, manual classification of Internet objects is an inadequate method of classification. Similarly, the use of unweighted or unverified text filtering alone results in inadequate and often inaccurate classification of Internet objects. Given the growing availability of adult content on computer-readable media, there is also need for a method and device that can more accurately and efficiently identify adult content on computer readable media, and either filter or deny access to such adult content.

The present invention can also be used to classify Internet objects and/or objects stored on computer readable media based on other criteria besides adult content.

SUMMARY OF THE INVENTION

An object of the present invention is to address the limitations of the prior art with respect to classification methods and devices for Internet objects (i.e. anything that is downloaded or transmitted via the Internet, including but not limited to web pages, images, text documents, email messages, newsgroup postings, chat text, video, and audio). The present invention addresses the limitations of the prior art

content. In order to understand how the robot navigates the Internet, it is necessary to understand the differences in the levels of addressing required to identify Internet objects. At the highest level, Internet addresses are governed by unique "domain names", such as playboy.com. Once granted this domain name, the owners of the domain name can set up computers that post Internet content at the "website" located at ~~http://~~www.playboy.com/. This Internet content takes the form of individual web pages. An Internet object that contains links to individual web pages is known as a directory. An example of a directory is ~~http://~~www.playboy.com/sex/. Each individual web page has a unique address, known as a universal resource locator, or URL. An example of a URL is ~~http://~~www.playboy.com/sex/morsex.html. The robot works by visiting and classifying Internet objects at individual URLs.

The robot takes as input a list of URLs and visits each one of them. At each site, the robot performs the classification analysis set out in detail below. At each site, the robot determines whether to add any new URLs found to its visitation list. This analysis is based on the level of adult content found on the URL. The robot maintains a list of URLs already visited so that no URL is visited more than once.

The robot builds and maintains a database of URLs which contain Internet objects which exceed the tolerated threshold for adult content. The database can be referenced upon executing Internet access via any communications port in any type of computer. Upon execution of Internet access, the database of URLs with adult

The Descriptor Coefficient

Many Internet objects have identifiers embedded in the object, known as meta data. Meta data is information stored within the Internet object. It is not normally visible to the viewer of an Internet object, but is machine-readable, and provides information about the Internet object that can be used by computer applications. Examples of meta data include the computer language in which the Internet object is written, the author or creator of the Internet object, and keywords that describe the Internet object.

Internet content providers currently have the option of embedding voluntary rating labels in the meta data of their Internet objects. There exist several such labelling systems. All of these systems allow an analysis of the content provider's opinion on whether the object contains adult content. The emerging standard in labelling systems is the PICS label system, whereby providers of Internet content voluntarily rate their Internet objects in terms of the amount of adult content (namely nudity, sex, violence and profane language) contained in the Internet object. A complete description of the rating guide for PICS values can be found at: ~~http://~~www.icra.org/ratingsv01.html. In essence, however, the nudity, sex, violence and profane language in the content of each Internet object is rated on a scale of zero to four, with four representing the most adult content in each of the categories.

For example, the PICS label for the web page ~~http://~~www.playboy.com/index.html consists of the following single line of meta text:

> "<meta http-equiv="pics-label" content='(pics-1.1 "http://www.rsac.org/ratingsv01.html" l gen true comment "RSACi North America Server" for http://www.playboy.com on "1996.05.04T06:51-0800" r (n 4 s 3 v 0 l 4))'>".

The actual PICS rating is found in the very last portion of the PICS label, and it tells a reader that the content provider of the Playboy website has rated the level of nudity on the site as a four out of four , the level of sex on the site sex as a three out of four, the level of violence on the site as a zero out of four, and the level of profane language on the site language as a four out of four.

Figure 3 illustrates the calculation of the descriptor coefficient for a PICS code. As illustrated, the PICS code (302) of an Internet object is read by the device of the invention and the individual elements of the PICS code, known as labels, are parsed (304). The labels are weighted to reflect the relative acceptability of the adult content signalled by the various items rated by the PICS label. The weighting scheme can be determined automatically or determined by the user. The ratings of the labels are multiplied by their weighting (306) and an average PICS code rating is determined by adding the weighted values (308) and dividing by the number of

2. let images = null

3. let audio = null

4. for each video component, v, in the object do

4.1. add each frame of v to images

4.2. add (concatenate) the audio stream of v to audio

5. set videocoeff = the image coefficient of images / (total number of frames)

6. set audiocoeff = the audio coefficient of the extracted audio data

7. set videocoeff = videoweight * videocoeff + audioweight * the audiocoeff of audio

8. output videocoeff

The Plug-in Coefficient

As technology evolves, new types of Internet objects are introduced. Producers of Internet software, most notably Netscape and Microsoft, have responded to this trend by introducing plug-ins. Plug-ins are programs that operate with viewing software, in this case web browsers and mail readers, and allow users to view Internet object types for which viewers were not originally included in the software. In most cases, these plug-ins are provided by the organization that developed the Internet object type. Current examples include Adobe's ~~Acrobat~~ ACROBAT ~~Reader~~ READER and Macromedia's ~~Flash Player~~ FLASH PLAYER.

The invention contemplates using information about new Internet object types by allowing a third party to produce plug-ins that can compute type specific

Accordingly, for Internet objects that have not yet crossed the threshold of objectionable adult content, the robot performs one final analysis based on relations between Internet objects. There are two types of relations that the robot considers: "part-of" relations, and "links-to" relations. An Internet object "A" is said to be "part-of" another Internet object "B" if A is contained in B, in some natural sense. Examples include: files in a directory, directories on a machine, machines in an Internet domain, and messages in a newsgroup. Of course, "part-of" relationships are transitive. In other words, if A is a part of B and B is a part of C, then A is a part of C. An Internet object A is said to "link to" another Internet object B if A contains a reference to B. Examples include web pages that link to other web pages and email or newsgroup postings that are replies to previous postings.

The robot does not consider "linked-to-by" relations, as the target Internet object has no control over the types of sites that link to its site. A "linked-to-by" relation is one in which the Internet object being classified is identified as having been the subject of a link from an object that has been previously classified as containing adult content. For example, many pornographic sites contain links to Internet filtering companies such as ~~NetNanny~~ NETNANNY, in order to provide consumers who wish to censor such sites with the means to do so.

Figure 10 illustrates the calculation of the relational coefficient. First, the robot compiles a list of all of the other Internet objects that an Internet object under consideration links to (the "linked-to"objects) (1002), as well as a list of all of the

```
4.   if u appears to be an adult-
        content site then
5.   add any links to HTML pages found on u to Q
6.    end if
7. end while
```

The robot maintains a list of URLs already visited so that no URL is visited more than once. Line 4 of the above algorithm is implemented as follows [where the site (hostname) of URL is denoted by site (URL); the directory of the URL on the host is denoted by directory (URL); and the domain is denoted as domain (URL) For example, if URL were ~~http://~~www.scs.carleton.ca/~morin/friends.html, then site (URL) would be ~~http://~~www.scs.carleton.ca, directory (URL) would be ~~http://~~www.scs.carleton.ca/~morin/, and domain would be carleton.ca.]:

```
1. if the web page domain(URL)appears to be adult content then
2. Mark URL as adult content.
3. if the web page site(URL) appears to be adult content then
4.   mark URL as adult-content
5. else if the web page directory(URL)
   appears to be adult content then
6.   mark URL as adult content
7. else if the web page URL appears to
   be adult-content then
8.   mark URL as adult content
9. else mark URL as not adult content
```